Module 4 Assignment | Lab 5: Spark

Trang Tran

CPS, Northeastern University

ALY6110 | Data Management and Big Data

Professor Andrew Kinley

Aug 08, 2023

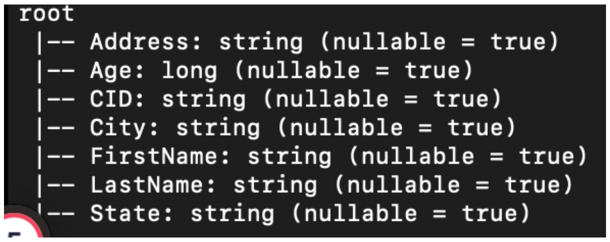
$ pyspark # (exit from pyspark with 'exit()')

1. customersDF = spark.read.json("SPRK/TutorialA/customers.json")

Read the json data from the "SPRK/TutorialA/customers.json" file and create a new dataframe named customersDF.

2. customersDF.printSchema()

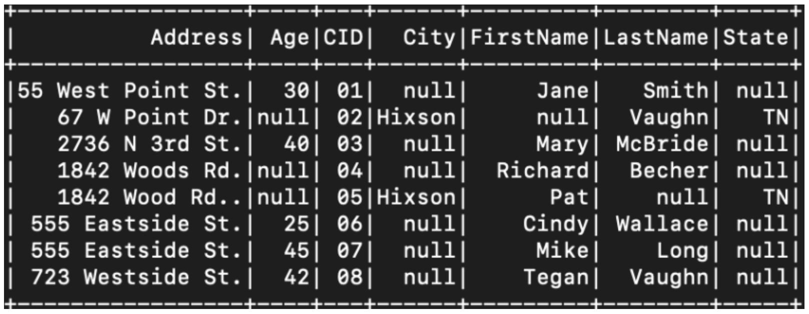
Print the schema of the ‘customersDF’ dataframe, including showing the data type of each column.



```
root
 |-- Address: string (nullable = true)
 |-- Age: long (nullable = true)
 |-- CID: string (nullable = true)
 |-- City: string (nullable = true)
 |-- FirstName: string (nullable = true)
 |-- LastName: string (nullable = true)
 |-- State: string (nullable = true)
```

3. customersDF.show()

Show the first few rows of the ‘customersDF’ dataframe, providing a preview of the data.



```
+-----------------+----+---+------+---------+--------+-----+
|          Address| Age|CID|  City|FirstName|LastName|State|
+-----------------+----+---+------+---------+--------+-----+
|55 West Point St.|  30| 01|  null|     Jane|   Smith| null|
|   67 W Point Dr.|null| 02|Hixson|     null|  Vaughn|   TN|
|   2736 N 3rd St.|  40| 03|  null|     Mary| McBride| null|
|   1842 Woods Rd.|null| 04|  null|  Richard|  Becher| null|
|   1842 Wood Rd..|null| 05|Hixson|      Pat|    null|   TN|
| 555 Eastside St.|  25| 06|  null|    Cindy| Wallace| null|
| 555 Eastside St.|  45| 07|  null|     Mike|    Long| null|
| 723 Westside St.|  42| 08|  null|    Tegan|  Vaughn| null|
+-----------------+----+---+------+---------+--------+-----+
```

4. First3DF = customersDF.take(3)

Create a new dataframe called ‘First3DF’ containing the first three rows of the ‘customersDF’ dataframe.

5. nameAgeDF = customersDF.select("FirstName","LastName", "Age")

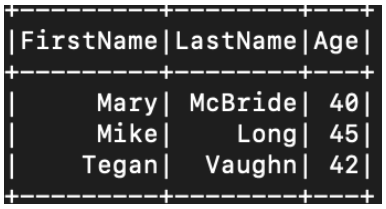
Create a new dataframe ‘nameAgeDF’ that contains only the "FirstName," "LastName," and "Age" columns from the ‘customersDF’ dataframe.

6. nameAgeOver30 = nameAgeDF.where("age>30")

Create a new dataframe named ‘nameAgeOver30’ by filtering the ‘nameAgeDF’ dataframe with only the rows where its "Age" column is greater than 30.

7. nameAgeOver30.show()

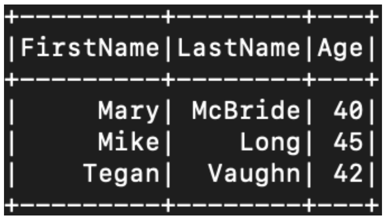
Show the first few rows of the ‘nameAgeOver30’ dataframe.



```
+---------+--------+---+
|FirstName|LastName|Age|
+---------+--------+---+
|     Mary| McBride| 40|
|     Mike|    Long| 45|
|    Tegan|  Vaughn| 42|
+---------+--------+---+
```

8. customersDF.select("FirstName", "LastName", "Age").where("Age>30").show()

This line is a combination of the 3 previous code lines. Do the selection of columns ("FirstName," "LastName," "Age") and filtering based on the "Age" column with values greater than 30, then displays a few rows of the result. The result below is exactly the same as the above.



```
+---------+--------+---+
|FirstName|LastName|Age|
+---------+--------+---+
|     Mary| McBride| 40|
|     Mike|    Long| 45|
|    Tegan|  Vaughn| 42|
+---------+--------+---+
```

9. myCustList = [["Mary","Luthans"],["Joyce","Lari"],["Kathi","Burge"]]

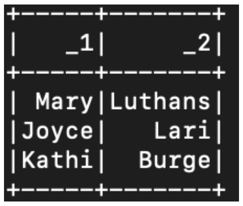
Create a list named ‘myCustList’, containing lists of customer data.

10. myCustDF = spark.createDataFrame(myCustList)

A dataframe named ‘myCustDF’ is created from the myCustList using the spark.createDataFrame() function.

11. myCustDF.show()

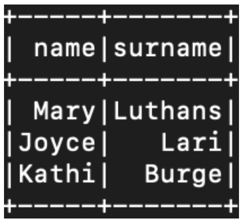
Show the first few rows of the ‘myCustDF’ dataframe. It consists of 2 columns and 3 rows.



```
+-----+-------+
|   _1|     _2|
+-----+-------+
| Mary|Luthans|
|Joyce|   Lari|
|Kathi|  Burge|
+-----+-------+
```

12. myCustDF.select("_1", "_2").withColumnRenamed ("_1","name").withColumnRenamed ("_2","surname").show()

Select the 2 columns from ‘myCustDF’ dataframe and rename them into ‘name’ and ‘surname’ columns, then display the first few rows of the result.



```
+-----+-------+
| name|surname|
+-----+-------+
| Mary|Luthans|
|Joyce|   Lari|
|Kathi|  Burge|
+-----+-------+
```

13. AddressDF = customersDF.select("FirstName","LastName", "Address")

Create a new dataframe named ‘AddressDF’ by selecting the "FirstName," "LastName," and "Address" columns from the ‘customersDF’ dataframe.

14. AddressDF.write.save("address")

Write and save the content of the ‘AddressDF’ dataframe to the "address" directory.

15. AddressDF.write.json("SPRK/TutorialB/address_list")

Write and save the contents of the ‘AddressDF’ dataframe to the "SPRK/TutorialB/address_list" directory in json format.

16. customersDF = spark.read.format("csv").option("header", "true").load("SPRK/TutorialB/customer.csv")

Read the CSV data format loading from the "SPRK/TutorialB/customers.csv" file into a dataframe named ‘customersDF’. The option("header", "true") indicates that the CSV file contains headers.

17. customersDF.select(customersDF["Age"])

18. customersDF.select("Age")

19. customersDF.select(customersDF.Age) // the extra command for Python

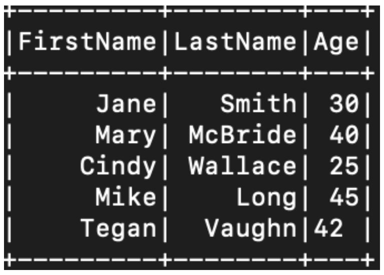
Lines 17, 18, and 19 show different ways to select the ‘Age’ column from the ‘customersDF’ dataframe.

20. ageDF = customersDF.select("FirstName", "LastName", "Age").where(customersDF.Age.isNotNull())

Create a new dataframe named ‘ageDF’ by selecting only the "FirstName," "LastName," and "Age" columns from ‘customersDF’, and filtering only the rows where the "Age" column is not null values.

21. ageDF.show()

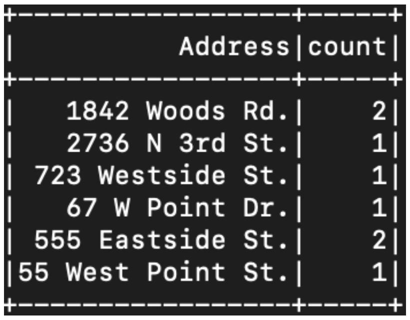
Display the first few rows of the ‘ageDF’ dataframe.



```
+---------+--------+---+
|FirstName|LastName|Age|
+---------+--------+---+
|     Jane|   Smith| 30|
|     Mary| McBride| 40|
|    Cindy| Wallace| 25|
|     Mike|    Long| 45|
|    Tegan|  Vaughn|42 |
+---------+--------+---+
```

22. customersDF.groupBy("Address").count().show()

Group the ‘customersDF’ dataframe by the "Address" column, count the frequency of each unique address, then show the address values and their corresponding count as follows.



```
+----------------+-----+
|         Address|count|
+----------------+-----+
|   1842 Woods Rd.|    2|
|   2736 N 3rd St.|    1|
| 723 Westside St.|    1|
|   67 W Point Dr.|    1|
| 555 Eastside St.|    2|
|55 West Point St.|    1|
+----------------+-----+
```